Infrastructure of an Institution Repository

CHRIS HELMS
GEORGIA TECH LIBRARY

GKR Symposium
August 9, 2012

Establishing a Service

- GKR will provide IR hosting and meta-searching services using open source software, along with IR-related services that include metadata and content submission, digital preservation, rights management, partner training, and content digitization.
- GKR will initially host 6 new DSpace repositories

Institutional Repository Hosting

Albany State University
POTENTIAL. REALIZED.

Georgia Health Sciences University

Georgia Institute of Technology

GALILEO

Georgia Gwinnett COLLEGE

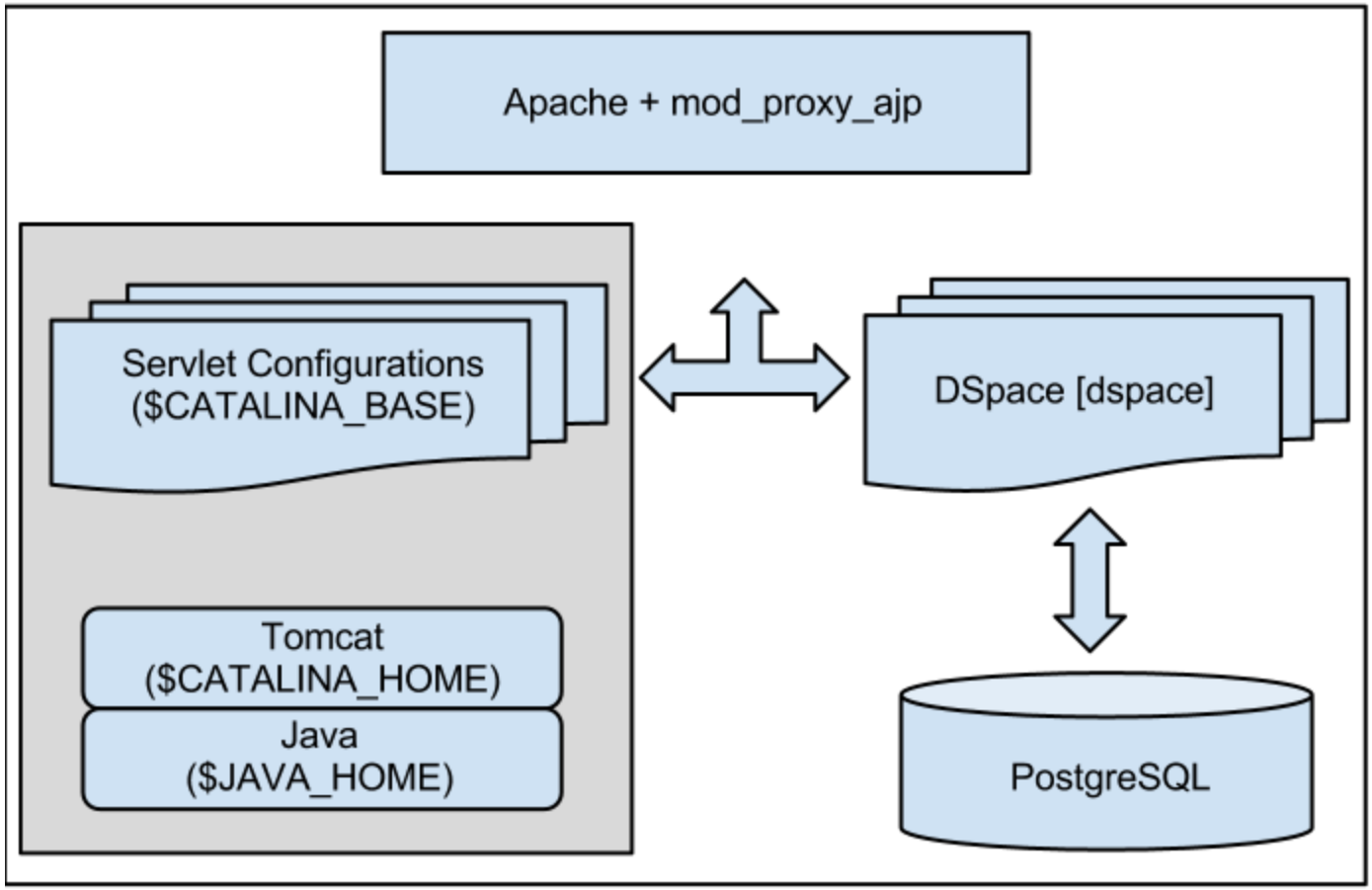
Goals

- Provide methods to implement multiple instances of DSpace that are independent of one another
  - Supporting multiple developers
  - Different patch sets, versions, or functionality if required
  - Discrete, easily identifiable assets
  - Branding
- Offer ways to tune each individual instance based on performance requirements and load

Physical Hardware

- Running on two PowerEdge R710
  - Primary (production)
  - Secondary (development, redundancy)
- 3TB (RAID) of usable storage space
- 24Gb Memory
- 2 Sockets, Multiple Cores
- Each physical machine would be able to sustain 6 instances of DSpace, one instance of PostgreSQL, and the mapping tool (Perl/CGI, Apache)

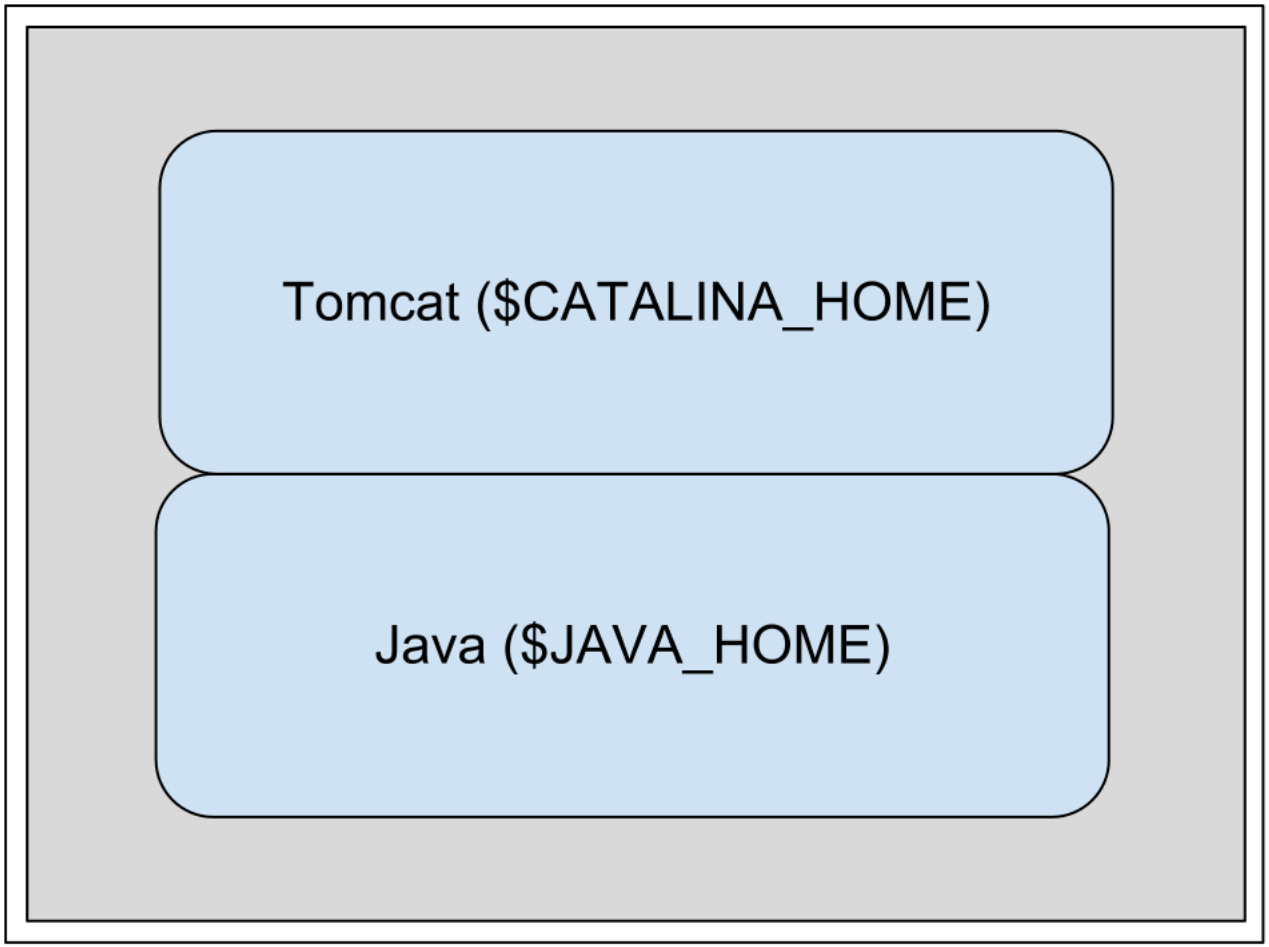
Technology Stack



Users, Ownership, & Groups

- Create one or more users for your DSpace projects
  - $ sudo useradd –c “DSpace GKR” dspace_gkr
  - $ sudo useradd –c “DSpace Proj1” dspace_proj1
- This user will also own and run a Tomcat instance
  - One user to multiple projects
  - Multiple users to multiple projects
- The created user’s home directory contains [dspace-source]
  - /home/dspace_gkr/dspace-1.8.2-src-release

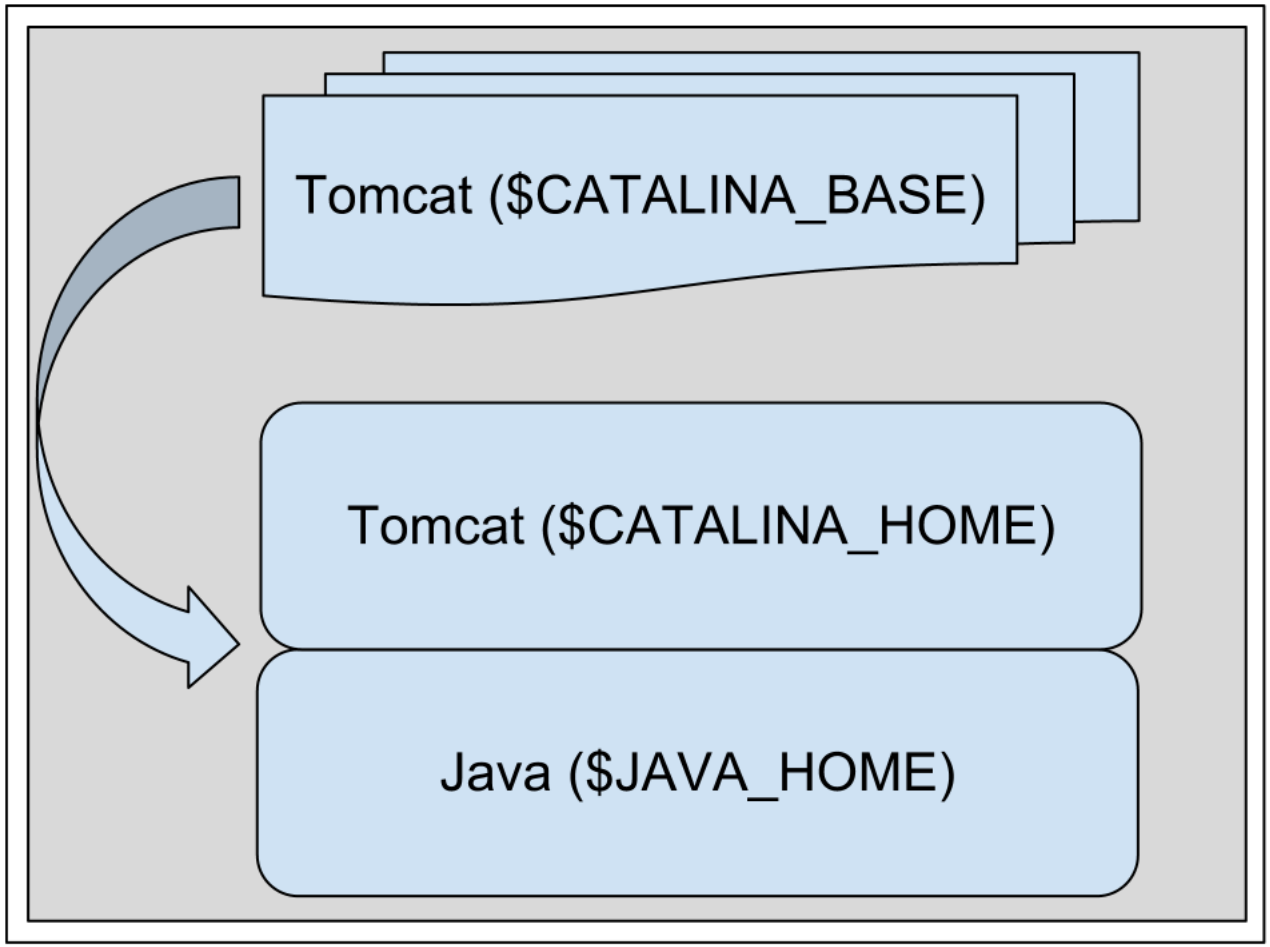
$HOME



Java and Catalina Home

- $JAVA_HOME
  - Java installation
  - Java (JRE or JDK)
  - /usr/java/latest or /usr/java/jdk1.6.0_32
  - JAVA_HOME="/usr/java/latest"
- $CATALINA_HOME
  - The location of our Tomcat installation
  - Binary unzip to /usr/local/apache-tomcat-6.0.35
  - ln –s /usr/local/apache-tomcat-6.0.35 /usr/local/tomcat
  - CATALINA_HOME="/usr/local/tomcat"

Tomcat Base Directories



What is CATALINA_BASE?

- The Tomcat reference directory for each Java servlet engine
  - Referred to as $CATALINA_BASE
  - Example “template” directory:
    - Setup the initial template directory in /var/www/tomcat_root
    - $ cd /var/www/tomcat_root
    - $ mkdir –p template/{bin,conf,logs,temp,work}
    - $ cp –R $CATALINA_HOME/conf/* template/conf/

Tomcat/DSpace Start | Stop

- Create a “control” script for your template
- Place the script within bin/
  - bin/template-control
- It will setup the running environment ($CATALINA_HOME, $CATALINA_BASE) for your instance of DSpace
- This script may be linked to boot time runlevel scripts (/etc/init.d, init)

template-control

```
#!/bin/bash
#
# Startup file for Tomcat: TEMPLATE
#
# chkconfig: 235 99 15
# description: Tomcat startup script for TEMPLATE
#
# Replace "template" with the name of your instance

# Set Tomcat/Java environment.
export JAVA_HOME="/usr/java/default"
JAVA_OPTS="-server -d64 -Xmx400M -Xms400M -Dfile.encoding=UTF-8"
JAVA_OPTS="$JAVA_OPTS -XX:PermSize=128M -XX:MaxPermSize=192M"
export JAVA_OPTS
export CATALINA_HOME="/usr/local/tomcat"
export CATALINA_BASE="/var/www/tomcat_root/TEMPLATE"
export CATALINA_PID="/var/www/tomcat_root/TEMPLATE/TEMPLATE.pid"
export PATH="/usr/java/default/bin:$PATH"

  [ -f $CATALINA_HOME/bin/startup.sh ] || exit 0
  [ -f $CATALINA_HOME/bin/shutdown.sh ] || exit 0

  # Set "run as" user and UID as needed for your
  # application
  RUNAS=TEMPLATE
  DSPACE_UID="505"
```

CATALINA_BASE :: Template Directory

```
/var/www/tomcat_root/template
|-- bin
|     |--template-control
|-- conf
|     |-- catalina.policy
|     |-- catalina.properties
|     |-- context.xml
|     |-- logging.properties
|     |-- server.xml
|     |-- tomcat-users.xml
|     |-- web.xml
|-- logs
|-- temp
|-- work
```

From here, create a copy of your template directory for each [dspace] install.

- $ cp –R template app1
- $ chown –R dspace_app1 app1

Template to app

- Rename your control-script, update the base variables
- app1/bin/app1-control

```
11  # Set Tomcat/Java environment.
12  export JAVA_HOME="/usr/java/default"
13  JAVA_OPTS="-server -d64 -Xmx400M -Xms400M -Dfile.encoding=UTF-8"
14  JAVA_OPTS="$JAVA_OPTS -XX:PermSize=128M -XX:MaxPermSize=192M"
15  export JAVA_OPTS
16  export CATALINA_HOME="/usr/local/tomcat"
17  export CATALINA_BASE="/var/www/tomcat_root/app1"
18  export CATALINA_PID="/var/www/tomcat_root/app1/app1.pid"
19  export PATH="/usr/java/default/bin:$PATH"
```

Assign unused ports

Application: I	Application: II	Application: III
Shutdown Port: 8005	Shutdown Port: 8009	Shutdown Port: 8017
AJP Port: 8006	AJP Port: 8010	AJP Port: 8018
HTTP Port: 8007	HTTP Port: 8011	HTTP Port: 8019
HDL HTTP: 8008	HDL HTTP: 8012	HDL HTTP: 8020
HDL TCP/UDP: 2874	HDL TCP/UDP: 2875	HDL TCP/UDP: 2876

Be sure to keep a list of port to application assignments

- /var/www/tomcat_root/PORT_ASSIGNMENTS

server.xml

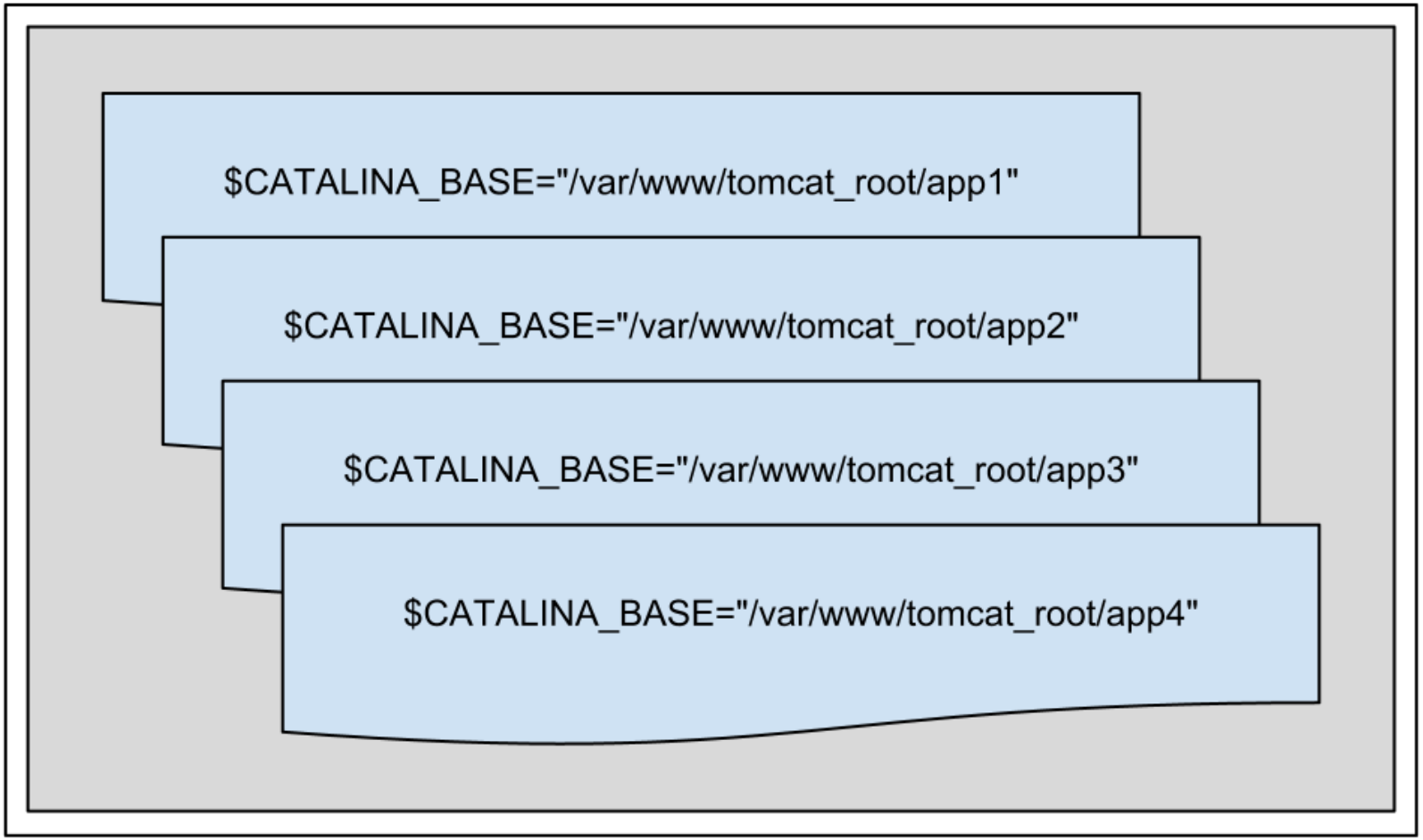
- Edit server.xml
  - $CATALINA_BASE/app1/conf/server.xml
  - 3 unique port numbers
  - Modify appBase and docBase to point to your [dspace] install directory

```
1  <?xml version='1.0' encoding='utf-8'?>
2
3  <Server port="8017" shutdown="SHUTDOWN">
4
```

server.xml continued

```
20  <Service name="Catalina">
21    <Connector port="8019" protocol="HTTP/1.1"
22               connectionTimeout="20000"
23               redirectPort="8443" URIEncoding="UTF-8" />
24    <!--
25    <Connector port="8443" protocol="HTTP/1.1" SSLEnabled="true"
26               maxThreads="150" scheme="https" secure="true"
27               clientAuth="false" sslProtocol="TLS" />
28    -->
29    <Connector port="8018" protocol="AJP/1.3" URIEncoding="UTF-8"
30        maxThreads="50" minSpareThreads="10" maxSpareThreads="25"
31        tomcatAuthentication="false" />
32
33    <Engine name="Catalina" defaultHost="localhost">
34
35      <Host name="localhost"  appBase="/var/www/dspace_root/dspace_app1/webapps"
36            unpackWARs="false" autoDeploy="true"
37            xmlValidation="false" xmlNamespaceAware="false">
38
39        <Context path="/" docBase="/var/www/dspace_root/dspace_app1/webapps/xmlui" debug="0"
40                 reloadable="true" cachingAllowed="false" allowLinking="true" />
```

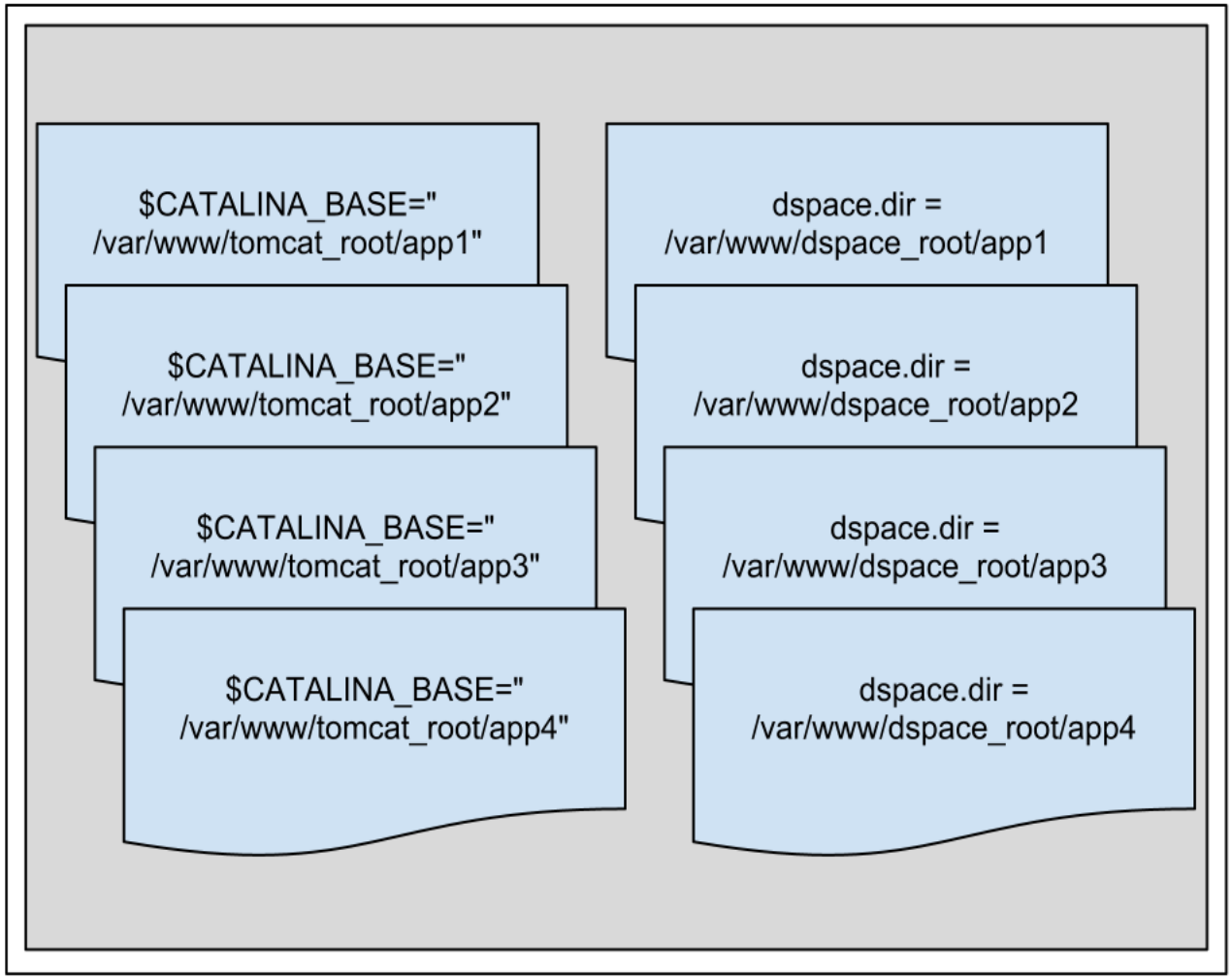
Results


$CATALINA_BASE="/var/www/tomcat_root/app1"
$CATALINA_BASE="/var/www/tomcat_root/app2"
$CATALINA_BASE="/var/www/tomcat_root/app3"
$CATALINA_BASE="/var/www/tomcat_root/app4"

DSpace Install

- Building from the DSpace documentation
  - [dspace-source] – Source directory from which you will install DSpace
    - /home/dspace_app1/dspace-1.8.2-src-release
  - [dspace] – The DSpace installation directory
    - /var/www/dspace_root/app1
  - [dspace]/webapps – The directory that contains your DSpace webapp
    - /var/www/dspace_root/app1/webapps (appBase)
- For each [dspace] directory you will need a corresponding $CATALINA_BASE directory

BASE to [dspace]

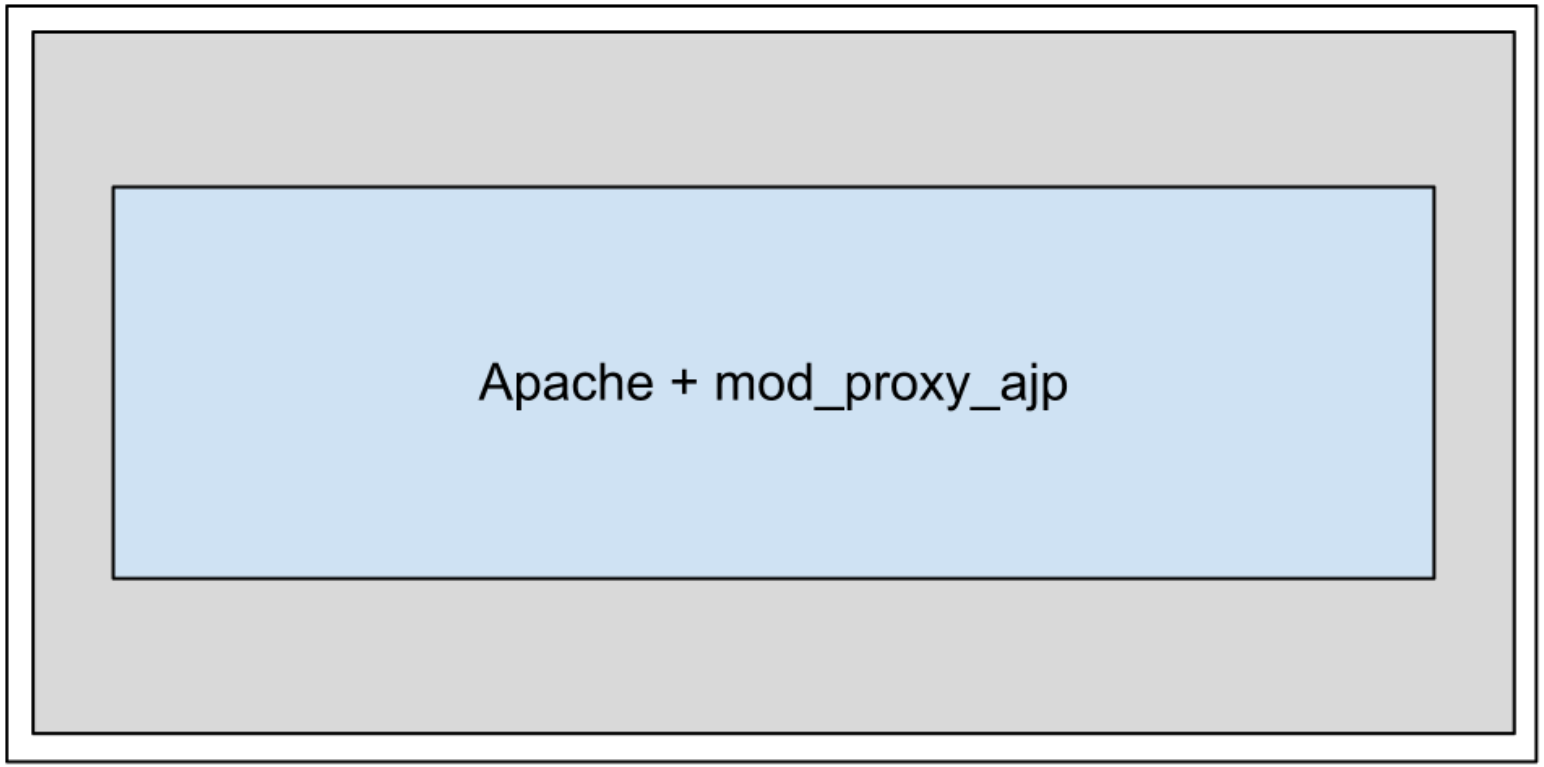


$CATALINA_BASE <-> [dspace]

```
/var/www/tomcat_root/app1
|-- bin
|     |--app1-control
|-- conf
|     |-- catalina.policy
|     |-- catalina.properties
|     |-- context.xml
|     |-- logging.properties
|     |-- server.xml
|     |-- tomcat-users.xml
|     |-- web.xml
|-- logs
|-- temp
|-- work
```

```
/var/www/dspace_root/app1
|-- assetstore
|-- bin
|-- config
|-- etc
|-- exports
|-- handle-server
|-- lib
|-- log
|-- reports
|-- search
|-- solr
|-- upload
|-- webapps
```

Apache



What was listening on that port?

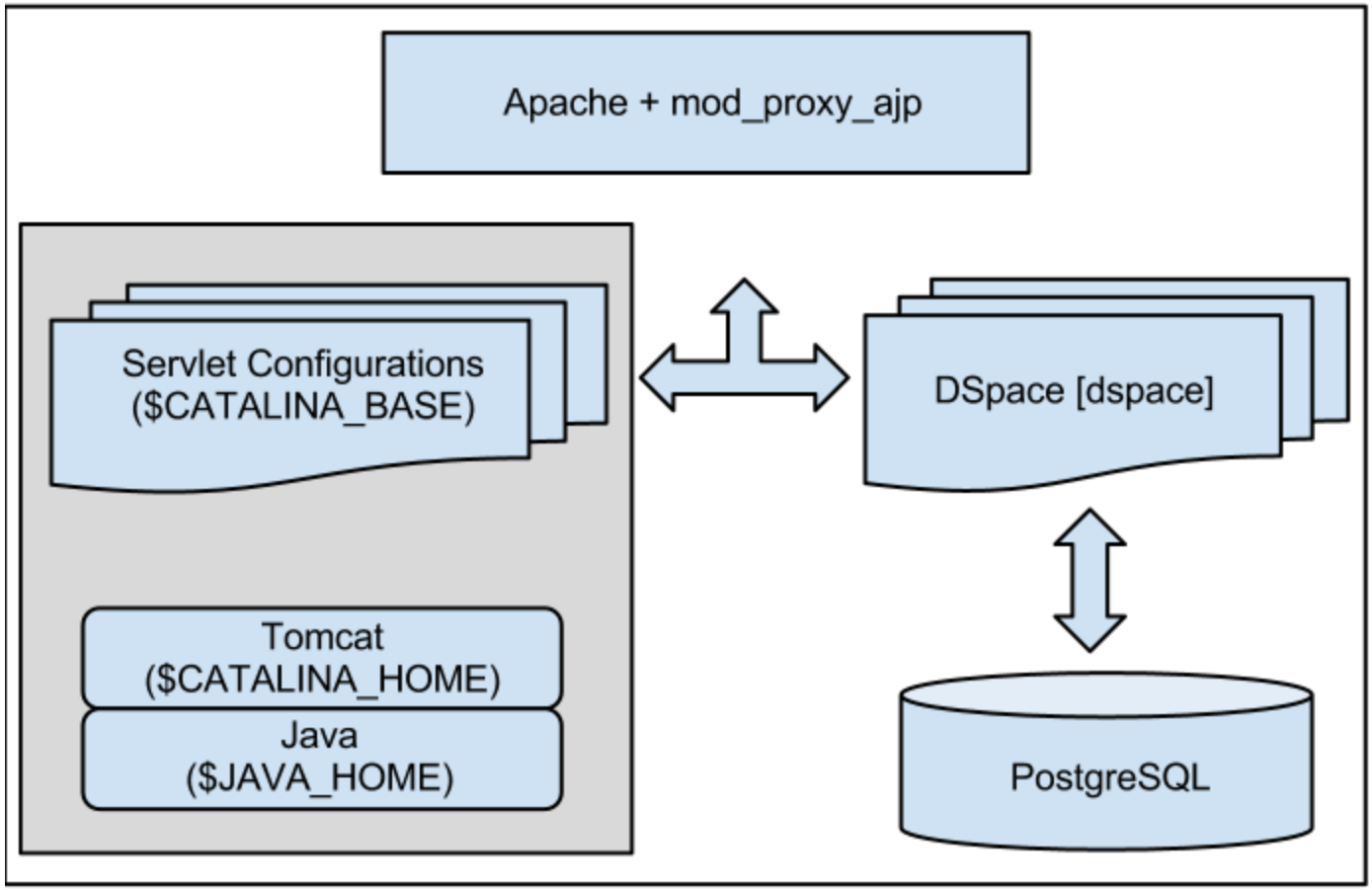
Application: I	Application: II	Application: III
Shutdown Port: 8005	Shutdown Port: 8009	Shutdown Port: 8013
AJP Port: 8006	AJP Port: 8010	AJP Port: 8014
HTTP Port: 8007	HTTP Port: 8011	HTTP Port: 8015
HDL HTTP: 8008	HDL HTTP: 8012	HDL HTTP: 8016
HDL TCP/UDP: 2874	HDL TCP/UDP: 2875	HDL TCP/UDP: 2876

/var/www/tomcat_root/PORT_ASSIGNMENTS

Apache virtual host

- ServerName www.gaknowledge.org
  - LoadModule proxy_ajp_module modules/mod_proxy_ajp.so
  - ProxyPass / ajp://localhost:8014/
  - ProxyPassReverse / ajp://localhost:8014/
- Alternative to mod_proxy_ajp
  - Tomcat Connectors – mod_jk
  - http://tomcat.apache.org/download-connectors.cgi

Technology Stack



Snapshots and Backups

- Nightly file system snapshots
  - Rsync – Fast incremental file transfer
    - http://rsync.samba.org/
    - MD5 checksum
    - Secure transfer via SSH
- Nightly database dumps
  - pg_dump -co
  - pg_dumpall –co
- Backup Target
  - HSM Archival Storage System
  - Controlled via SAM-QFS (ZFS filesystem)
  - Disk cache, Tape cache

Which directories should I backup?

```
/var/www/tomcat_root/app1
|-- bin
|     |--app1-control
|-- conf
|     |-- catalina.policy
|     |-- catalina.properties
|     |-- context.xml
|     |-- logging.properties
|     |-- server.xml
|     |-- tomcat-users.xml
|     |-- web.xml
|-- logs
|-- temp
|-- work
```

```
/var/www/dspace_root/app1
|-- assetstore
|-- bin
|-- config
|-- etc
|-- exports
|-- handle-server
|-- lib
|-- log
|-- reports
|-- search
|-- solr
|-- upload
|-- webapps
```

Fin

Questions? Comments?

Chris Helms
chris.helms@library.gatech.edu
Georgia Tech Library